Configuration and Optimization of the Hewlett-Packard SureStore E Disk Array XP256 In the MPE/iX Environment

By Stephen F. Macsisak
Hewlett-Packard Company
19447 Pruneridge Avenue.MS 47UA
Cupertino, CA 95014
408 447 5851
Fax 408 447 4278

Paper ID#-455

Configuration and Optimization of the Hewlett-Packard SureStore E Disk Array XP256 In the MPE/iX Environment

Stephen F. Macsisak
Hewlett-Packard Company

Paper ID#-455

The HP SureStore E Disk Array XP256 is a large multi-terabyte, scaleable, enterprise-capable highly available disk array. The XP256 supports many platforms including the HP e3000. Since the first installation on a HP e3000 in September 1999, many enterprise class HP e3000 customers have taken advantage of the XP256 to decrease their vulnerability to disk failure and to improve performance. In addition, through use of tools like HP SureStore E Business Copy XP, HP e3000 users have increased their backup flexibility and reduced their vulnerability to system hardware problems.

The XP256 uses 15GB, 37GB and 48GB high-speed disks in RAID 1 and RAID 5 configurations with built-in hot spares to eliminate the likely hood of a disk failure causing down time. The XP256 supports up to nine terra bytes of space in an easily expanded environment to provide support for multiple systems/platforms with the same disk system. The XP256 supports up to 16 GB of cache memory in the disk controller cabinet; in some environments, this provides for read from/write to cache hits for increased performance.

In an HP e3000 environment with the XP256, increased batch performance, faster transaction manager checkpoints and reasonable levels of cache read hits are seen. The only caveat is that MPE/iX is highly optimized for On Line Transaction Processing (OLTP) and can be spindle-limited in small configurations. Unlike other enterprise disk array systems, the XP256 has a high performance RAID 5 implementation. Most HP e3000 XP256 users have some or all of the their disks configured as RAID 5 resulting in increased storage capacity without reducing performance. In RAID 1 configurations, one-half of the physical storage is used as a backup for the other half. In RAID 5 configurations, one-fourth of the physical storage is used to maintain enough information so a single disk failure will not lose data.

In the first few XP256 installations, MPE/iX systems were required to have logical device one on a disk configured as RAID 1. This restriction has been lifted and some users are deploying XP256 system with all the disks in RAID 5 configurations. The only recommendation is that the configuration (actually any HP e3000 system) should include enough physical spindles to keep the system from having any disk volumes with over fifty percent disk I/O utilization. Currently the HP e3000 only supports FWD (Fast/Wide Differential) SCSI-2 connections. In the typical HP e3000 OLTP environment, the average disk I/O is about 8KB in length and a single FWD SCSI-2 controller can easily support 250 I/Os per second while only consuming ten percent of the bandwidth of the SCSI bus (the upper limit is 20MB per second).

The disk industry is continually increasing the capacity of disks without making the same rate of improvements in disk speeds. Now you can buy 18 GB FWD SCSI-2 (soon 32 or 72GB) disks and use them on your HP e3000 (it seems like only yesterday MPE/iX went from supporting 2GB SE SCSI-2 disks to 4GB FWD SCSI-2 disks). Disk RPMS (revolutions per minute) have increased, but going from a 5400-RPM disk to 10000-RPM disk does not double disk performance: it only increases disk performance by forty to sixty per cent. Replacing four 4GB 5400 FWD SCSI-2 disks with one 18GB 10000 RPM disk, you might experience worse performance if disk I/O demand was more than the single spindle could handle. Of course, it really depends on your current I/O rates and whether you have one job serially reading a disk or one hundred OLTP users trying to read all the disks on the system at the same time.

A very rough rule of thumb: to have enough physical spindles to have good or better performance than you currently have, you probably need to increase your disk space by two to three times when going to a XP256 if you currently have 100 GB of JBOD (just a bunch of disks) and have a high disk I/O rate. You can’t just configure your new XP256 with the half the number of physical spindles you currently have and expect the read cache hits to make up the difference. Batch performance will probably increase on an XP256 because the disk system reads further ahead than even MPE/iX does when it detects batch access.

In looking at MPE/iX OLTP environments, which have large cached disk systems, a typical system, achieves about a 25-45 percent cache read hit rate. A reasonable approach would be to assume one-third of the I/Os in your current environment are converted to cache hits, which complete in a millisecond or two. The other two-thirds of your I/Os will have to come from disk just like in a JBOD environment.

Another good approach to follow in configuring a disk environment is to strive for twenty-five percent utilization across all disks. In the disk I/O context, disk utilization is defined as the number of I/Os per second issued divided by the maximum I/Os per second achievable by a disk. For example, if the average access time of a disk were 15 milliseconds, the disk would be capable of about 66 I/Os per second at 100 percent utilization. If the disk I/O rate to the same disk was actually 33 I/Os per second, consider the disk to be fifty percent utilized. The same disk would be twenty-five percent utilized if the I/O rate was 16.5 I/Os per second. When a disk is accessed at higher utilization levels, it’s likely that one or more I/Os are already queued for the disk, which causes queuing delays. Another rule of thumb is that if disk utilization is fifty percent, then disk I/Os take twice as long as at twenty-five percent because of queuing delay. At seventy-percent utilization, a disk I/O takes four times as long because of queuing delays. At eighty percent utilization, an I/O takes six times as long and response time starts experiencing exponential delays. The graph in figure 1 on the next page illustrates the phenomena.

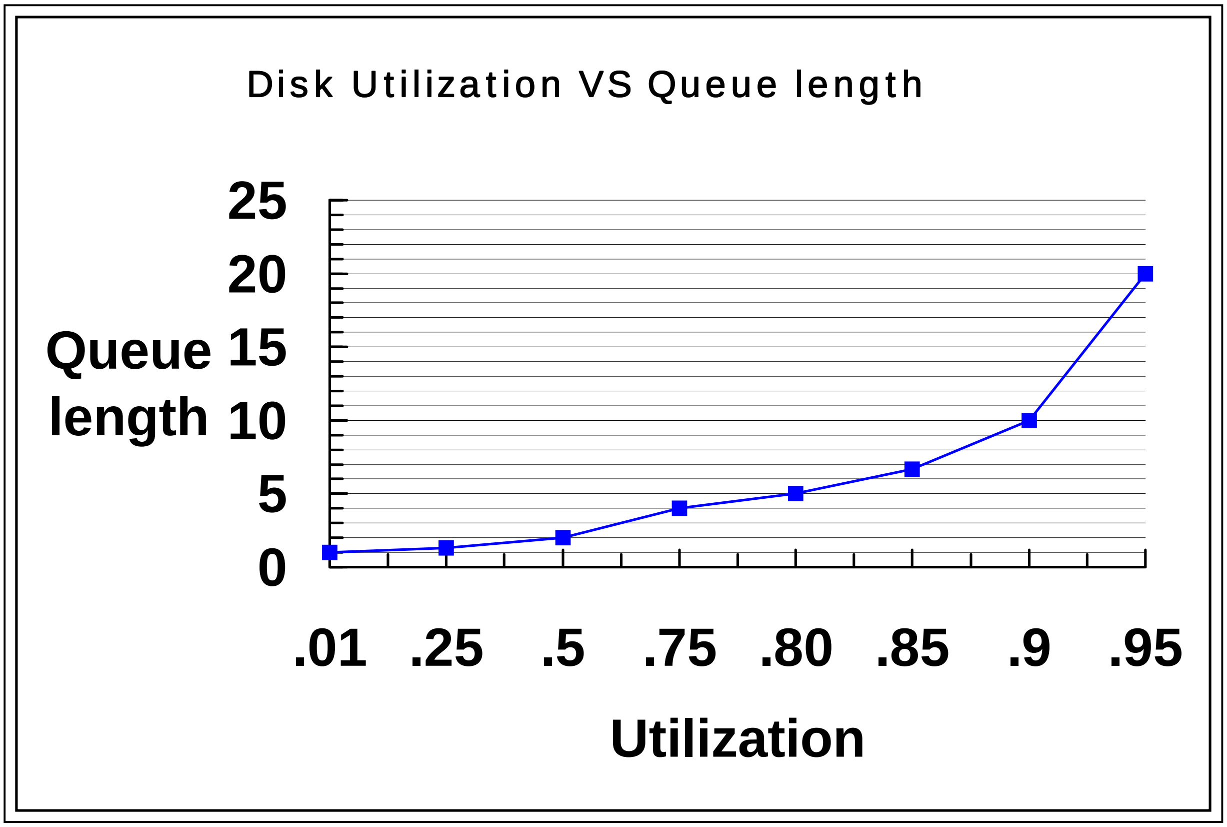

Figure 1

After about fifty percent disk utilization, as Disk utilization goes up, the queue length starts increasing much faster. The basic rule is that as Disk Utilization goes up, the probability increases that a Disk I/O is already in front of your request.

The 15 GB 12000 RPM disks used in the XP256 have an average access time of about nine milliseconds. In calculating the number of physical spindles, use a number that takes into account the latency of the whole system; for estimation purposes, use about 15 milliseconds per disk I/O.

Now you have the assumptions necessary to estimate the number of spindles required for performance in an XP256 MPE/iX system. Take a hypothetical situation where you need to do 500 Disk I/Os per second and have excellent OLTP performance combined with the High Availability capabilities of the XP256.

Assumptions used in the calculation:

25 percent disk utilization
33 percent hit rate
15 millisecond average access time

A single spindle can do about 66 Random I/Os per second and if you want to achieve 25 percent disk utilization, use 66/4 or about 16.5 I/Os per second per spindle as a target. Next, assume that two-thirds of your I/Os are misses and result in physical I/Os to the disk. You need to have about twenty physical spindles ((500*2/3)/16.5) to achieve good performance with these parameters. The XP256 system with this configuration would have about have about 225 GB (20*.75*15) of usable space if utilizing a Raid 5 setup. A comparable unprotected JBOD configuration with 4GB FWD SCSI-2 disks would have about 30 physical spindles and about 120 GB of usable space. As with any general performance recommendation, "it depends". If doing an actual configuration, you need to work with a storage specialist.

Adding more cache in the XP256 will help with read hits on many different platforms. MPE/iX is already optimized for OLTP and makes good use of main memory, tending to reduce the read hit percentage. The cache does help in other ways by reducing Transaction Manager (XM) contention. During a checkpoint, MPE/iX writes out data from memory to the disk volume set that has previously been logged to the XM journal file located on the master volume of the volume set. The large amount of disk I/O issued during a checkpoint is normally transparent to the online user but in some very intensive modification environments, disk contention during checkpoints can cause an increase in response time. The XP256 with at least four GB of cache can take the data as fast as MPE/iX can write to the XP without causing disk contention. If you have an application that uses TurboIMAGE block on commits at the end of transactions or is using XDBBEGIN/XDBEND calls, then the ability of the XP256 to do immediate completions of disk writes can improve performance without sacrificing transactional integrity.

One MPE/iX, the file system exists in memory and on disk. At any point in time (except just after an XM recovery operation), the disk environment is in an unknown state and depends on the fact that all accesses go through memory which may contain changes unwritten to disk. If the system were to fail, then XM recovery would undo (if incomplete) or post (if complete) system or user transaction in the XM journal file that had not been written to disk by a checkpoint. Some customers have purchased HP SureStore E Business Copy XP to make local copies of data for backup, testing or application development. In addition, to make Business Copy work well in the MPE/iX environment, the data needs to be on an MPE/iX user volume set and a VSCLOSE command must be issued before the "link" to the copy is broken. The VSCLOSE command assures the data on the user volume set is in a known state (i.e., all data is posted and no files are open on the user volumes of the set). Users may have to exit the application or logoff to cause all files on the set to be closable. Once the link is broken, the Business Copy disks can be mounted on another HP e3000. The same restrictions apply to HP SureStore E Continuous Access XP. If the "link" is broken without using the VSCLOSE command, the data on the BC disks is in the same state as MPE/iX disks are after a system abort: the disks will have to go through XM recovery when mounted.

The XP256 is a versatile disk system. Some MPE/iX users of the XP256 are considering using the XP as a fail-over mechanism. They will have a cold standby system. If the primary system is down for more than some minimum time for hardware problems, the site will boot another system off the disks connected to XP256 and “take over” all functions of the down system. This standby system enables IT management to make guarantees about the length of down time.

Other users are looking at a similar scenario but will have a hot standby secondary system. If the primary system is down, the disks from a user volume set can be mounted. The users will still have to reconnect to the new system but downtime can be minimized. Using an XP256 in this way allows flexibility in planning OS and hardware upgrades. At a slack time, the users can log off, the drives can be switched using the XP256 to another HPe300 and remounted. Online users can then reconnect to the secondary system while the primary system is being upgraded. Commercial Systems Division is working on products to automate many of the tasks associated with these switching operations.

Currently there are many MPE/iX sites using the XP256. One of the first sites utilizing the product has over 2k users connected to four MPE/iX systems, all connected to one XP256. Another site has a large MPE/iX system sharing the XP256 with a large HP-UX system. At another site, the XP256 replaced a disk farm of 120 mirrored (240 total) MPE/iX disk volumes and was able to reduce the backlog the application experienced at times. Other sites are connecting the HP e3000 to an XP256 that will be shared with WINDOWS NT.

Using the XP256 to “harden” the disk systems of an HP e3000 has proven to be a very effective way to increase the already legendary reliability of MPE/iX and the HP e3000.